Learning properties of Support Vector Machines

Arnaud Buhot, Mirta B. Gordon*
Département de Recherche Fondamentale sur la Matière Condensée,
CEA/Grenoble, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France
(February 16, 1998)

In this article, we study the typical learning properties of the recently proposed Support Vectors Machines. The generalization error on linearly separable tasks, the capacity, the typical number of Support Vectors, the margin, and the robustness or noise tolerance of a class of Support Vector Machines are determined in the framework of Statistical Mechanics. The robustness is shown to be closely related to the generalization properties of these machines.

Support Vector Machines, recently proposed to solve the problem of learning classification tasks from examples, have aroused a great deal of interest due to the simplicity of their implementation, and to their remarkable performances on difficult tasks [1,2]. Classification of data is a very general problem, as many real-life applications, like pattern recognition, medical diagnosis, etc., may be cast as classification tasks. In the last few years, much work has been done to understand how high-performance learning may be achieved, mainly within the paradigm of neural networks. These are systems composed of interconnected neurons, which are two-state units like spins. The neuron's state is determined, like in magnets, by the sign of the weighted sum of its inputs, which acts as an external field, and of the states of its neighbors. Learning with neural networks means determining their connectivity and the weights of the connections. The aim is to classify correctly not only the examples, or training patterns, but also new data, as we expect that the learning system will be able to generalize. A single neuron connected to its inputs, the *simple perceptron* (SP), is the elementary neural network. It separates the input patterns in two classes by a hyperplane orthogonal to a vector whose components are the connection weights. Thus, the SP can learn without errors only Linearly Separable (LS) tasks. Most classification problems are not LS, requiring learning machines with more degrees of freedom. However, the relationship between the machine's complexity, its learning capacity and its generalization ability is still an open problem. Feedforward layered networks, the *multilayered perceptrons*, are the most popular learning machines. Their architecture is usually found through a trial and error procedure, in which the weights are determined with Backpropagation [3], a learning algorithm that performs a gradient descent on a cost function. Its main drawback is that it usually gets trapped in metastable states. Growth heuristics that avoid using Backpropagation have also been proposed [4].

Support Vector Machines (SVM) [1,2,5] are an alternative solution to the learning problem, whose *typical* properties have not been studied theoretically yet. The idea underlying SVM is to map the patterns from the input space to a new space, the *feature-space*, through a non-linear transformation chosen *a priori*. Provided that the dimension of the feature-space is large enough, the image of the training patterns will be LS, *i.e.* learnable by a SP. It is well known that, if the training set is LS, there is an infinite number of error-free separating hyperplanes. Among them, the *Maximal Stability Perceptron* (MSP) has weights that maximize the distance of the patterns closest to it. The SVM weight vector is that of the MSP in feature-space. The patterns closest to the separating hyperplane are called *Support Vectors* (SV) ; their distance to the hyperplane is the *maximal stability* or SV-margin. The important point is that the SV determine uniquely the MSP. Their number is proportional to the number of training patterns, and *not* to the dimension of the feature-space (which may be huge). Thus, increasing the feature-space dimension does not necessarily increase the number of parameters to be learned, a fact that makes the SVM very attractive for applications. For example, in the problem of digit recognition [1], the input space of dimension 256 needs to be mapped onto a space of dimension $256^7 \sim 10^{16}$, but the number of parameters to be determined is as low as 422. However, in spite of the high performance reached by SVMs in realistic problems [2], a theoretical understanding of their properties is still lacking.

We consider, within the framework of Statistical Mechanics, SVMs defined by particular families of mappings between the input-space and the feature-space. We address several important questions about these machines. The generalization error in the particular case of learning a LS task is shown to decrease slower than that of a SP (in input-space) as a function of the number of training patterns. The capacity increases proportionally to the dimension of the feature-space. The number of SV and the SV-margin present interesting scaling with the number of features. The probability of misclassification of training patterns corrupted after learning is shown to be a decreasing function of the SV-margin. This property, that we call robustness or noise-tolerance, may account for the good generalization performance of SVMs in applications.

We assume that we are given a training set of P in-

dependent N-dimensional vectors, the *training patterns* $\{\boldsymbol{\xi}^\mu\}_{\mu=1,\cdots,P}$, and their corresponding classes $\tau^\mu = \pm 1$. The patterns are supposed to be drawn with a probability density $P(\boldsymbol{\xi}) = (2\pi)^{-N/2} \exp\left(-\boldsymbol{\xi}^2/2\right)$, and the classes τ are given by an unknown function $\tau(\boldsymbol{\xi})$ called supervisor or teacher.

We focus on SVMs defined by a nonlinear transformation Φ that maps the N-dimensional input space to a $(k+1)N$-dimensional feature-space through

$$\boldsymbol{\xi} \to \Phi(\boldsymbol{\xi}) = \{\boldsymbol{\xi}, \phi(\lambda_1)\boldsymbol{\xi}, \cdots, \phi(\lambda_k)\boldsymbol{\xi}\}, \tag{1}$$

where the λ_i are functions of $\boldsymbol{\xi}$. The components $\phi(\lambda_i)\boldsymbol{\xi}$ $(i = 1, \cdots, k)$ are the *new features* that hopefully should make the task linearly separable in feature-space.

In the following we consider odd functions ϕ, and $\lambda_i = \boldsymbol{\xi}\cdot\mathbf{B}_i$ where $\{\mathbf{B}_i\}_{i=1,\cdots,k}$ is a set of k unitary orthogonal vectors ($\mathbf{B}_i\cdot\mathbf{B}_j = \delta_{ij}$). With this choice, the new features are uncorrelated. For example, the k first generators $\{e_1, e_2, \cdots, e_k\}$ of the input space ($e_1 = (1, 0, \cdots, 0)$, $e_2 = (0, 1, 0, \cdots, 0), \cdots$) are one possible realization of the $\mathbf{B}_i$. In the thermodynamic limit considered below, any set of k randomly selected normalized vectors $\mathbf{B}_i$ satisfies the orthogonality constraint with probability one. The functions $\phi(\lambda) = \text{sign}(\lambda)$ and $\phi(\lambda) = \lambda$ are of particular interest. If $k = N$, a SVM using the latter can implement all the possible discriminating surfaces of second order in input space. More complicated transformations Φ, equivalent to higher order surfaces, may be considered (for examples, see [1]).

The output of the SVM to a pattern $\boldsymbol{\xi}$ is $\sigma = \text{sign}\left(\mathbf{J}\cdot\Phi(\boldsymbol{\xi})\right)$, where $\mathbf{J} = \{\mathbf{J}_0, \mathbf{J}_1, \cdots, \mathbf{J}_k\}$ is a $(k+1)N$-dimensional vector. Hereafter we consider normalized weights, $\mathbf{J}\cdot\mathbf{J} = (k+1)N$ without any lack of generality, but we *do not* impose any constraint to the normalization of each N-dimensional vector $\mathbf{J}_i$. The stability of a training pattern $\boldsymbol{\xi}^\mu$ of class τ^μ in feature-space is

$$\gamma^\mu = \frac{\tau^\mu\ \mathbf{J}\cdot\Phi(\boldsymbol{\xi}^\mu)}{\sqrt{(k+1)N}}. \tag{2}$$

Geometrically, $|\gamma^\mu|$ is the distance of the image $\Phi(\boldsymbol{\xi}^\mu)$ of pattern $\boldsymbol{\xi}^\mu$ to the hyperplane orthogonal to $\mathbf{J}$. The aim of learning is to determine a vector $\mathbf{J}$ such that $\sigma^\mu = \tau^\mu$, or equivalently $\gamma^\mu > 0$, for all μ. Any vector $\mathbf{J}$ that meets these learning conditions separates linearly, in the feature-space, the image Φ of patterns with output $+1$ from those with output -1. Due to the non-linearity of Φ, this separation is not linear in input space. More general SVMs, that use a Kernel $K(\mathbf{J}, \Phi(\boldsymbol{\xi}))$ instead of the inner product in Eq.(2), have been proposed [1], but we restrict to the inner product in the following.

The SV-margin is

$$\kappa_{\max}(\mathbf{J}^*) = \max_{\mathbf{J}} \inf_{\mu} \gamma^\mu, \tag{3}$$

where $\mathbf{J}^*$, the MSP weight vector in feature-space, is a linear combination of the SV [1,6], $\mathbf{J}^* = \sum_{\mu\in SV} x^\mu \tau^\mu \Phi(\boldsymbol{\xi}^\mu)$. The x^μ are positive parameters to be determined by the learning algorithm, which has to determine also the number of SV. Generally, this number is small compared with the feature-space dimension, a fact that allows to increase the latter considerably without increasing dramatically the number of parameters to be determined. The SVM in input-space ($k = 0$) or *linear SVM* is the usual MSP, whose properties have extensively been studied (see [14] and references therein).

We obtain the generic properties of the SVM through the by now standard replica approach [15]. Results are obtained in the thermodynamic limit, in which the input space dimension and the number of training patterns go to infinity ($N \to +\infty$, $P \to +\infty$) keeping the reduced number of patterns $\alpha = P/N$ constant. In this limit, the SVM properties are independent of the training set. The appropriate cost function is $E(\mathbf{J}, \mathcal{L}_\alpha, \kappa) = \sum_\mu \Theta(\kappa - \gamma^\mu)$, where Θ is the Heaviside function and $\mathcal{L}_\alpha$ represents the training set. It counts the number of training patterns that have a stability smaller than κ in feature-space. The largest value of κ that satisfies $E(\mathbf{J}^*, \mathcal{L}_\alpha, \kappa) = 0$ is the SV-margin. The weight vector $\mathbf{J}^*$ defines the SVM. Its generic properties are determined by the free energy

$$f = \lim_{N\to+\infty} \lim_{\beta\to+\infty} -\frac{1}{\beta N} \langle \ln Z \rangle, \tag{4}$$

where $Z = \int dP(\mathbf{J}) \exp\left(-\beta E(\mathbf{J}, \mathcal{L}_\alpha, \kappa)\right)$ is the partition function, $dP(\mathbf{J}) = d\mathbf{J}\ \delta((k+1)N - \mathbf{J}\cdot\mathbf{J})$ and β is an inverse temperature. In Eq. (4), the bracket stands for the average over all the possible training sets $\mathcal{L}_\alpha$ at given α. If the problem is LS, then $f = 0$ for $\kappa \geq 0$, meaning that error-free learning is possible. But in general, the probability of error-free learning vanishes beyond some value of κ. The maximal value of κ for which $f = 0$ is the *typical* value of $\kappa_{\max}(k, \alpha)$. The free energy is calculated using the replica trick $\langle \ln Z\rangle = \lim_{n\to 0} \ln \langle Z^n \rangle / n$.

We consider first the case of a teacher that is a SP in input space, of (unknown) N-dimensional normalized weigths $\mathbf{K}$ ($\mathbf{K}\cdot\mathbf{K} = 1$). Thus, the classes of the training patterns $\boldsymbol{\xi}^\mu$ are $\tau^\mu = \text{sign}\left(\mathbf{K}\cdot\boldsymbol{\xi}^\mu\right)$. In this case, an error-free solution exists for all α, and we are interested in the generalization error $\epsilon_g(k, \alpha)$, which is the probability that the trained SVM misclassifies a new pattern $\boldsymbol{\xi}$. Clearly, we do not expect that a SVM with $k > 0$ will perform well on this task, as it corresponds to a case where the *a priori* selected feature-space is too complex. However, this may well be the case in real applications. We begin by considering this LS problem mainly because other properties considered below, like the capacity and robustness, can easily be deduced by disregarding, or setting to zero, some of the order parameters introduced here. These are,

$$R^a = \frac{\mathbf{J}_0^a\cdot\mathbf{K}}{\sqrt{\mathbf{J}_0^a\cdot\mathbf{J}_0^a}}, \tag{5a}$$

$$v_i^a = \frac{\mathbf{J}_i^a\cdot\mathbf{J}_i^a}{N}, \tag{5b}$$

$$c_i^{ab} = \lim_{\beta \to +\infty} \beta \frac{(\mathbf{J}_i^a - \mathbf{J}_i^b)^2}{2N} \qquad (a \neq b), \tag{5c}$$

for $i = 0, \cdots, k$. $\mathbf{J}^a$ and $\mathbf{J}^b$ are the weight vectors of replicas a and b. The cross-overlaps $\mathbf{J}_i^a \cdot \mathbf{J}_j^b$ $(i \neq j)$, and $\mathbf{K} \cdot \mathbf{B}_i$ may be neglected for $k \ll N$, as they are of order $1/\sqrt{N}$. The parameters c_i^{ab} are a generalization of the parameter $x^{ab} = \lim_{\beta \to +\infty} \beta(1 - q^{ab})$ in [8]. In fact, Gardner and Derrida [8,11] considered a single perceptron $(k = 0)$ with normalized weights $\mathbf{J}_0$ $(\mathbf{J}_0 \cdot \mathbf{J}_0 = N)$, so that $(\mathbf{J}_0^a - \mathbf{J}_0^b)^2/2N = 1 - \mathbf{J}_0^a \cdot \mathbf{J}_0^b/N = 1 - q^{ab}$ in their notations. We assume replica symmetry, *i.e.* $R^a = R$, $v_i^a = v_i$, $c_i^{ab} = c_i$ for all a, b. The parameter R represents trivially the overlap between the first N components of vector $\mathbf{J}$ with the teacher $\mathbf{K}$. The overlap between $\mathbf{J}_i$ and $\mathbf{K}$ may be neglected for $i \geq 1$, since for odd functions ϕ and uncorrelated vectors $\mathbf{B}_i$ the new features are uncorrelated. If ϕ were even, this would not be the case. The parameters v_i are proportional to the norm of the $\mathbf{J}_i$. The sense of the parameters c_i is more involved. They reflect how fast the fluctuations of $\mathbf{J}_i$ around the minimum of the cost function decrease as the temperature vanishes $(\beta \to +\infty)$. In the case of a degenerate continuum of minima, these fluctuations decrease too slowly, and the c_i diverge. This is the case for $\kappa < \kappa_{\max}$.

A symmetry between the k vectors $\mathbf{J}_i$, $i \geq 1$, due to the invariance with respect to permutations of the $\mathbf{B}_i$, together with the fact that the $\mathbf{B}_i$ are uncorrelated with $\mathbf{K}$, allows to take $v_i = v_1$ and $c_i = c_1$ for $i \geq 1$. Introducing $\tilde{v}_1 = v_1/v_0$, where v_0 is determined by the normalization condition $\mathbf{J} \cdot \mathbf{J}/N = k + 1 = v_0 + k\tilde{v}_1 v_0$, $\tilde{c}_1 = c_1/c_0$ and $\tilde{c}_0 = c_0/(1 + k)$, the free energy is $f(k, \alpha, \kappa) = \max_{\tilde{v}_1, \tilde{c}_1, \tilde{c}_0} \min_R g(k, \alpha, \kappa; \tilde{v}_1, \tilde{c}_1, \tilde{c}_0, R)$, with

$$\begin{aligned} g(k, \alpha, \kappa; \tilde{v}_1, \tilde{c}_1, \tilde{c}_0, R) &= -\tfrac{\tilde{c}_1\,(1-R^2) + k\tilde{v}_1}{2\,\tilde{c}_0\,\tilde{c}_1\,(1+k\tilde{v}_1)} \\ + \ \tfrac{\alpha}{\tilde{c}_0} \int D\lambda_1 \cdots \int D\lambda_k \int_{\kappa a - b}^{\kappa a} Dy\ &\tfrac{(\kappa - y/a)^2}{e} H\left(-\tfrac{yR}{\sqrt{e}}\right) \\ + \ 2\,\alpha \int D\lambda_1 \cdots \int D\lambda_k \int_{-\infty}^{\kappa a - b} Dy H&\left(-\tfrac{yR}{\sqrt{e}}\right). \end{aligned} \tag{6}$$

$Dy = dy \exp(-y^2/2)/\sqrt{2\pi}$, $H(x) = \int_x^{+\infty} Dy$, and a, b, e stand for

$$a = \sqrt{\frac{1 + k\tilde{v}_1}{e + R^2}}, \tag{7a}$$

$$b = a\left[\tilde{c}_0\left(1 + \tilde{c}_1 \sum_{i=1}^{k} \phi^2(\lambda_i)\right)\right]^{1/2}, \tag{7b}$$

$$e = 1 - R^2 + \tilde{v}_1 \sum_{i=1}^{k} \phi^2(\lambda_i). \tag{7c}$$

The generalization error $\epsilon_g(k, \alpha)$ writes

$$\epsilon_g(k, \alpha) = \frac{1}{\pi} \int D\lambda_1 \cdots \int D\lambda_k \ \arccos\left(\frac{R}{\sqrt{e + R^2}}\right), \tag{8}$$

where R and e extremize $g(k, \alpha, \kappa; \tilde{v}_1, \tilde{c}_1, \tilde{c}_0, R)$. In particular, the maximal stability $\kappa_{\max}(k, \alpha)$ is the largest value of κ that satisfies $\tilde{c}_0(\alpha, \kappa) = +\infty$ since f is non zero for finite values of $\tilde{c}_0$.

If $\phi(\lambda) = \text{sign}(\lambda)$, the extremization of (6) with respect to $\tilde{v}_1$ and $\tilde{c}_1$ gives $\tilde{v}_1 = 1 - R^2$ and $\tilde{c}_1 = 1$. Notice that for $R = 1$ (which corresponds to $\alpha = \infty$), $\tilde{v}_1 = 0$ (thus, $v_1 = 0$) as expected: the new features are irrelevant because the task is LS. The fact that $\tilde{c}_1 = 1$ means that the fluctuations of $\mathbf{J}_0$ and $\mathbf{J}_i$, $i \geq 1$, have the same behaviour in the limit $\beta \to \infty$ despite the fact that their norms are different $(\tilde{v}_1 \neq 1)$. After introduction of these values for $\tilde{v}_1$ and $\tilde{c}_1$ in (6), we obtain $g(k, \alpha, \kappa; \tilde{c}_0, R) = g\left(0, \alpha/(k+1), \kappa; \tilde{c}_0, R/\sqrt{1 + k(1 - R^2)}\right)$, where the right hand side term corresponds to a SP trained with a training set of reduced size $\alpha/(k+1)$ having an overlap $R/\sqrt{1 + k(1 - R^2)}$ with the teacher. After introducing these values of the order parameters in (7c) and (8), we obtain $\epsilon_g(k, \alpha) = \epsilon_g(0, \alpha/(k+1))$. As expected, the generalization error of the SVM with $k > 0$ on a LS task is larger than the one of the linear SVM. This is due to an entropic effect, as the SVM's phase space grows with k whereas the size of the space of functions considered, limited to the LS ones, remains the same. For large α, the generalization error vanishes as $0.5005\,(k + 1)/\alpha$, to be compared to the linear SVM that has $\epsilon_g \sim 0.5005/\alpha$ [14].

From the above scaling, the SV-margin and the number of SV follow from the maximal stability $\kappa_{max}(0, \alpha)$ and the distribution of stabilities $\rho(0, \alpha; \gamma)$ of the MSP in input space [14]. We obtain $\rho(k, \alpha; \gamma) = (\sqrt{2/\pi})\rho_1(k, \alpha)\,\Theta\left(\gamma - \kappa_{max}(k, \alpha)\right) + \rho_0(k, \alpha)\,\delta\left(\gamma - \kappa_{max}(k, \alpha)\right)$ where $\rho_1(k, \alpha) = H[-\gamma/\tan(\pi\epsilon_g(k, \alpha))]\exp(-\gamma^2/2)$ and $\rho_0(k, \alpha)$, the typical fraction of training patterns that belong to the SV, is such that $\rho(k, \alpha; \gamma)$ integrates to one. For $\alpha \ll 1$, the SV-margin is $\kappa_{max}(k, \alpha) \sim \sqrt{(k + 1)/\alpha}$ and $\rho_0(k, \alpha) \sim 1 - \sqrt{2\alpha/\pi(k + 1)}\exp -(k + 1)/2\alpha$, meaning that in that limit almost all the training patterns are SV. For $\alpha \to \infty$, $\kappa_{max}(k, \alpha) \sim 0.226\sqrt{2\pi}(k + 1)/\alpha$, and $\rho_0(k, \alpha) \sim 0.952(k + 1)/\alpha$, *i.e.* the typical number of SV is slightly smaller than the feature-space dimension. Solutions for other functions ϕ are more complicated, and we were not able to find a closed expression of $\epsilon_g(k, \alpha)$ for all α. The function ϕ that gives the smallest generalization error at given k, at least for small α, is $\phi(\lambda) = \text{sign}(\lambda)$. But the fact that the generalization error increases with k is a general property, independent of the function ϕ.

We turn now to the more interesting problem of the capacity, defined as the typical number of dichotomies that the SVM may implement, a quantity closely related to the VC dimension of the learning machine [7]. We consider training sets where the patterns' classes are given by a random teacher, that selects outputs $+1$ and -1 with the same probability $1/2$. In this case, the order parameters are (5b) and (5c). The free energy is $f(k, \alpha, \kappa) = \max_{\tilde{v}_1, \tilde{c}_1, \tilde{c}_0} g(k, \alpha, \kappa; \tilde{v}_1, \tilde{c}_1, \tilde{c}_0)$ where $g(k, \alpha, \kappa; \tilde{v}_1, \tilde{c}_1, \tilde{c}_0)$ is obtained from (6) and (7) by setting $R = 0$.

The capacity $\alpha_c(k)$, the largest reduced number of patterns that the machine can learn without errors, corresponds to a vanishing SV-margin, *i.e.* $\kappa_{\max}(k, \alpha_c(k)) = 0$. In this case, the extremae of $g(k, \alpha, 0; \tilde{v}_1, \tilde{c}_1, \tilde{c}_0)$ correspond to $\tilde{c}_0(\alpha, \kappa) = +\infty$ and $\tilde{v}_1 = \tilde{c}_1$ for all the possible functions ϕ. This result means that the capacity is $\alpha_c = 2(k+1)$, independently of ϕ, provided that the new features are uncorrelated. This result generalizes to other feature-spaces the value deduced by Cover [12] through a geometrical approach that Mitchison and Durbin [13] generalized to the case of quadratic separating surfaces. Notice that the latter corresponds to a SVM with $\phi(\lambda) = \lambda$ and $k = N$. The capacity of SVMs is smaller than the one of multilayered perceptrons with one hidden layer of $k+1$ neurons, which have the same number of degrees of freedom. For example, the capacity of the parity machine scales like $k \ln k$, and that of the committee machine like $k\sqrt{\ln k}$, for large k [9,10].

It turns out that in the case $\phi(\lambda) = \mathrm{sign}(\lambda)$, the maximal stability $\kappa_{\max}(k, \alpha)$ scales trivially with k. The order parameters are $\tilde{v}_1 = \tilde{c}_1 = 1$ so that $g(k, \alpha, \kappa; \tilde{c}_0) = g(0, \alpha/(k+1), \kappa; \tilde{c}_0)$, where the RHS corresponds to a SP of margin κ in input space. The maximal stability is thus $\kappa_{\max}(k, \alpha) = \kappa_{\max}(0, \alpha/(k+1))$. From [14] we deduce that for $\alpha \ll 1$, $\kappa_{\max}(k, \alpha) \sim \sqrt{(k+1)/\alpha}$, and for $\alpha \to \alpha_c^-$, $\kappa_{\max}(k, \alpha) \sim \sqrt{\pi/8}\,(2(k+1)/\alpha - 1)$. If $\phi(\lambda) = \lambda$, the property $\kappa_{\max}(k, \alpha) \sim \kappa_{\max}(0, \alpha/k)$ is correct for $\alpha \ll k$. As $\kappa_{\max}(0, \alpha)$ is a concave decreasing function of α [8], including new features may result in a large increase of the SV-margin.

In most classification problems we expect that similar patterns belong to the same class. In that case, having a large SV-margin may be benefical for the generalization performance. In particular, if slightly corrupted versions of the training patterns are presented to the trained SVM, its output should not change. We consider a SVM that achieved error-free learning with a SV-margin $\kappa_{\max} > 0$. We assume that the training patterns are corrupted, after the learning process, through $\boldsymbol{\xi}^\mu \to \boldsymbol{\xi}^\mu + \boldsymbol{\eta}^\mu$, where $\boldsymbol{\eta}^\mu$ are randomly distributed vectors with probability distribution:

$$P(\boldsymbol{\eta}) = (2\pi\Delta)^{-N/2} \exp\left(-\boldsymbol{\eta}^2/2\Delta^2\right). \quad (9)$$

We are interested in the classification error of the SVM on the corrupted patterns, defined as $\epsilon_t(k, \kappa, \Delta) = \sum_\mu (\sigma^\mu(\Delta) - \tau^\mu)^2/(4P)$ where τ^μ is the original pattern's class and $\sigma^\mu(\Delta)$ the SVM's output to the corrupted pattern. The dependance on α is implicitly included through $\kappa = \kappa_{\max}(k, \alpha)$. ϵ_t caracterizes the robustness of the SVM with respect to a small pattern's corruption ($\Delta \ll 1$). Input vectors close to a training pattern will be given its same class with probability $1 - \epsilon_t(\kappa, \Delta)$.

In the case of the linear-SVM (the SP), a straightforward calculation gives

$$\epsilon_t(0, \kappa, \Delta) = H(-\kappa)H(\kappa/\Delta) + \int_\kappa^{+\infty} Dz H(z/\Delta). \quad (10)$$

If the margin is $\kappa = 0$, one half of the training patterns have zero stability, and $\epsilon_t(0, 0, \Delta) > 1/4$. Thus, any small perturbation results in misclassifications. If $\kappa > 0$, then $\epsilon_t(0, \kappa, \Delta) \sim \exp(-\kappa^2/2\Delta^2)$ for small Δ. Consider next the general SVMs. If $\phi(\lambda) = \mathrm{sign}(\lambda)$ and $\kappa > 0$, $\epsilon_t(k, \kappa, \Delta) \sim \Delta$ for small Δ. In comparison with the SP, the robustness of SVM is poor. This is due to the discontinuity of the function ϕ, as a small perturbation of the input pattern may produce a strong perturbation on its stability. On the contrary, for continous functions ϕ, like $\phi(\lambda) = \lambda$, and small Δ, $\epsilon_t(k, \kappa, \Delta) \sim \exp(-h(k)\kappa/\Delta)$ where $h(k)$ is an increasing function of k. Thus, continous functions ϕ are preferable for improving the SVM's robustness or noise tolerance.

In conclusion, we presented the first study of the typical properties of a class of SVMs. We determined, as a function of the number of new features and the number of training patterns, the fraction of SV, the behaviour of the margin, the generalization error on a linearly separable task, the capacity and the probability of misclassification of training patterns slightly corrupted. Our results may explain why maximizing the margin is so important : the probabilty that the trained SVM will assign the same class to the corrupted as to the original training patterns is enhanced by large margins.

---

* and Centre National de la Recherche Scientifique.

[1] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, New York (1995).

[2] C. J. C. Burges, Preprint (1997).

[3] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, **323**, 533-536 (1986).

[4] J. M. Torres Moreno and M. B. Gordon, *Neural Computation*, **10**, 1017-1040 (1998)

[5] C. Cortes and V. Vapnik, *Machine Learning*, **20**, 273-297 (1995).

[6] U. Gerl and F. Krey, *J. Phys. I France*, **7**, 303-327 (1997).

[7] M. Opper, *Phys. Rev. Lett.*, **72**, 2113-2116 (1994).

[8] E. Gardner and B. Derrida, *J. Phys. A: Math. Gen.*, **21**, 271-284 (1988).

[9] R. Monasson and R. Zecchina, *Phys. Rev. Lett.*, **75**, 2432-2435 (1995).

[10] Y. Xiong,J.-H. Oh and C. Kwon, *Phys. Rev. E*, **56**, 4540-4544 (1997).

[11] E. Gardner, *J. Phys. A: Math. Gen.*, **21**, 257-270 (1988).

[12] T. M. Cover, *IEEE Trans. Elect. Comp.*, **14**, 326-334 (1965).

[13] G. J. Mitchison and R. M. Durbin, *Biol. Cybern.*, **60**, 345 (1989).

[14] M. B. Gordon and D. R. Grempel, *Europhys. Lett.*, **29**, 257-262 (1995).

[15] M. Opper and W. Kinzel, in *Models of Neural networks III*, E. Domany, J. L. van Hemmen and K. Shulten (Eds.), pp. 151-209 (Springer Verlag, New York, 1996).